Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-03 16:08:17
REFERENCE NO: 207

Author Names & Affiliations

- Kiran Keshav - Yale Center for Research Computing
- Andrew Sherman - Yale Center for Research Computing
- Robert Bjornson - Yale Center for Research Computing
- Sohrab Ismail-Beigi - Yale School of Engineering & Applied Science
- James Duncan - Yale School of Engineering & Applied Science
- Daisuke Nagai - Yale Department of Physics

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computer Science

Title of Submission

Yale Center for Research Computing - Response to NSF CI 2030 Request for Information

Abstract (maximum ~200 words).

A recent National Academies report stated: “Large-scale simulation and the accumulation and analysis of massive amounts of data are revolutionizing many areas of science and engineering research. Increased advanced computing capability has historically enabled new science, and many fields today rely on high-throughput computing for discovery.” At Yale, we see the emergence of huge quantities of data at multiple scales as a major driver of transformative scientific research. We highlight challenges in three particular disciplines: materials science, bioimaging, and astrophysics. To address these challenges, many innovations in CI will be crucial: better networks combining high speed with secure access and authentication; new, scalable storage architectures supporting better management, access, curation, and preservation; improved cybersecurity technologies supporting controlled data access essential to collaborative research; new CI funding models encouraging more flexible responses to infrastructure needs (especially storage); development of inter-institutional or regional communities of CI expertise; and enhanced cloud-based methodologies facilitating inter-institutional collaborations. We also discuss two other topics of great importance: CI workforce development and training, including creation of long-term career paths; and the need for direct connections between agency awards of research funding and guarantees of sufficient and appropriate CI to complete the research.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

According to a recent report from the National Academies of Sciences, Engineering, and Medicine: “Large-scale simulation and the accumulation and analysis of massive amounts of data are revolutionizing many areas of science and engineering research. Increased advanced computing capability has historically enabled new science, and many fields today rely on high-throughput computing for discovery. Modeling and simulation, the historical focus of high-performance computing, is a well-established peer of theory and experiment.”

At Yale, we see the emergence of huge quantities of data at multiple scales as a major driver of transformative scientific research in a number of disciplines. Below are three examples that highlight future cyberinfrastructure needs as a result of this data deluge.

In the design and engineering of “hard materials,” computation is being applied to ever more complex materials systems (i.e., more realistic materials models) involving multiple length scales, time scales, and dynamic and evolving materials. Aside from the sheer size and scale of such computations, a major research challenge confronting the entire field is to build reasonable and representative starting structures for the modeling. The problem is acute since experimental characterization is often unable to provide the sufficiently fine-grained structural information over large spatial domains needed to specify a (relatively) unambiguous structural model. In addition to the importance of enhancing characterization tools to routinely provide high resolution structural information on materials, computational materials scientists will routinely need to simulate a large number of possible realizations of the materials structure (all obeying whatever constraints are known from experiments) to build statistical models of how the material may behave. This will require novel approaches to building large classes of "reasonable" or "likely" models (terms whose meaning must be defined as part of the research endeavor) that are to be simulated in parallel and from which one should be able to draw useful conclusions. Looking further ahead, as advanced automated experimental characterization tools come on line (and some already exist in the form of specialized advanced x-ray or electron beams systems that produce 3D structural data on materials), the sheer volume and variety of data that they will produce will present a challenge. This will require computational advances coupled with high speed storage and on-the-fly computation that is co-located at the characterization facility. Only in this way will useful information be extracted from the experiments to help make computation and experiment go hand in hand and make synchronized progress.

Another important example is bioimaging, which provides a wealth of information that has a broad range of applications in basic research, and in clinical settings, including both disease diagnosis and the guidance and management of therapeutic interventions across many clinical disease domains, including neurodevelopmental disorders, cardiovascular disease and cancer. Information is typically acquired from a variety of imaging modalities and assembled post hoc, often without an a priori strategy of optimizing image acquisition and analysis to the diagnostic or therapeutic task at hand. In addition, image-based information can be derived at multiple scales, such as microscopic images from pathology or in vivo endoscopy to organ level imaging from computed tomography (CT) or magnetic resonance imaging (MRI). Going forward, it will be critical to integrate and optimize innovative image acquisition across multiple scales and multiple modalities with state-of-the art advances in computational analysis and machine learning for image-based diagnosis, therapy targeting and outcomes assessment. Ever more sophisticated computational and statistical models will be required to make sense of the data as well as to quantify results to guide subject-specific analysis/diagnosis and intervention, potentially even providing online feedback to the acquisition system.

Clearly, bioimaging is a field well-positioned to take advantage of the latest large-data-driven processing/analysis concepts based on many-layered neural network architectures (deep learning) for both image-derived biomarker assessment and prediction. However, currently, large image data repositories must be equipped with some sort of record keeping or normalization to account for variations in data acquisition, such as those across acquisition sites (i.e. hospital or imaging center) or across imaging-equipment manufacturer. In addition, patient/data confidentiality (i.e. HIPAA) issues are challenging and must be addressed. To handle all of this, new paradigms must be developed based on cloud computing concepts that can integrate information across multiple imaging scales (e.g. cellular imaging all the way to millimeter scale organ imaging), multiple energy sources/modalities (e.g. optical imaging using fluorescent dyes, MRI, PET, X-Ray, CT, ultrasound) and multiple physical locations (imaging centers) in order to assemble both training and test data for the algorithms of the future. To date, such cloud-based strategies and thinking have yet to be fully developed or thought out. High-throughput hierarchical data pipelines must be created that are sensitive to the bandwidth requirements of both processing algorithms and data transport. This may require, for instance, that computational kernels be mapped to appropriate processing hardware (e.g., GPUs) located close to the data banks holding the image information.

Finally, as a third specific example, we cite astrophysics. It is not possible to perform direct laboratory experiments for many astronomical phenomena due to the extreme time, distance, and energy scales involved. Advanced cyberinfrastructure must be used to create virtual universes and probe in detail the dynamics of complex astronomical phenomena such as the formation of galaxy clusters or the merging of two black holes. Almost every subfield of astronomy—from the study of origin, evolution and fate of planets, stars and galaxies to the large-scale structure (LSS) of the universe—has been impacted by continuing improvements in computational algorithms and hardware. This includes our ability to observe the sky, enabling massive surveys such as the upcoming LSST project, which will scan the visible sky every

night to unprecedented depth, generating petabytes of data in the process.
Cosmological simulations have become an essential tool for studying the formation and evolution of galaxies, galaxy clusters, and LSS of the Universe. Starting from well-defined initial conditions of the early universe, modern cosmological simulation codes can accurately follow how dark matter and normal matter, such as gas and stars, collapse to form galaxies and galaxy clusters and how these different components interact throughout a build-up of the LSS. Such simulation codes must resolve phenomena at multiple spatial and temporal scales, incorporating numerous coupled dynamical equations to address the effects of gravity and hydrodynamics, as well as sub-grid models that are beginning to address the formation of stars and black holes. In the future, the next steps in making astrophysical simulations increasingly realistic will require simulations to follow many more variables at each spatial location to contend with significant new complications associated with improved sub-grid models for the physics of star and galaxy formation, radiation hydrodynamics in non-axisymmetric environments, stiff and complex chemical or nuclear reaction networks, and multi-phase media including non-thermal components such as cosmic rays and magnetic fields. In order to validate the results of such complex simulations, it will be essential to compare results to the observational data from the LSST and other projects. In aggregate, these projects are causing the amount of astrophysical data to grow exponentially, leading to challenges both in storing it and in bringing the data and associated computations together. Very likely, some sort of public cloud-based solution will be required because the scale of the data will be beyond the capabilities of a single institution and because the computation and data must be co-located.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

In order to meet the challenges outlined in our response to Q1, a number of innovations in Cyberinfrastructure will be crucial.

Improved networks:

Obviously, as datasets explode in size, networks are increasingly the key bottleneck, both within and between institutions. Beyond simple speed improvements, we also need better ways to allow secure access without throttling speed, and improved global credentials for authentication between institutions.

Storage management, preservation, and access:

Current storage technologies are fundamentally based on monolithic posix-style file systems. This architecture is badly stretched at multi-petabyte scale, and will not scale much further. We need new storage architectures that are scalable to Exabyte scale, while remaining robust and easy to use. We also need powerful reporting and management systems for data, including access control, automated life cycle, tiering, metadata searching, content hashes, etc.

In addition, data produced for sponsored research should be safely archived in compliance with the terms of the funding agencies, and made available for future research. This requires a rethinking of how such data is cataloged, with unique identifiers, digital signatures, and data descriptors. It also requires a way to fund long term storage obligations that last long beyond the active term of a project.

Improvements to cyber security:

We need better ways to control access to data, allowing seamless access to collaborators across multiple institutions, including restricted data such as PHI, while preventing unauthorized access. Ideally sensitive data would be encrypted at rest.

Changes to funding model for CI:

Currently, most academic institutions are discouraged from “renting” computation cycles or storage to research projects, because such rentals are typically charged overhead, while capital purchases are not. This causes researchers to inappropriately make long term capital purchases of resources, which have a long lead time and are fixed, 5-8 year investments. This problem is particularly acute for storage, which must be purchased in very large units to be economical. Researchers would benefit from a model that allowed them more flexibility to buy or rent resources as needed.

Additional funding for CI and Application Specialists:

As high performance computing and big data spreads further from its traditional domains, there is a critical need for more application specialists who understand both the domain and computing. Naï¶¥ use of HPC can easily result in both wasted resources and incorrect r results. Experts in CI and applications are in high demand outside academia. Providing a mechanism by which experts in computing and applications can be stably supported would greatly improve the efficiency and quality of grant-funded research.

Support for regional CI specialist meetings:

Building inter-institutional communities among CI specialists is important, since the number at each institution is typically small. Large, yearly, general purpose meetings such as Supercomputing tend to be too large for this purpose. As an alternative, sponsoring frequent, short (one or two day), focused regional meetings would be a time-efficient and low-cost way to encourage cross-pollination and collaboration among CI specialists and researchers at nearby institutions.

Cloud-based methodologies:

Cloud computing's value to research is not primarily cost-driven. Rather, it is the promise it represents for huge, inter-institutional collaborative projects, to avoid data movement and replication by having a central repository for all data, to which computations can come. Unfortunately, the current cloud APIs are poorly suited to HPC, and are either too inflexible or too low level for the typical researcher to use efficiently. Cloud resources are also difficult to integrate with local resources.

In addition, federal agencies should seek to remove the administrative burden of using cloud resources with data which is subject to specific data use agreements and/or government regulations (e.g., HIPAA data).

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

In this section, we address two additional important topics.

Workforce Development and Training:

Surprisingly, despite the importance of advanced research computing and the vast amounts of money spent on computational resources, workforce development—the human element of computational science—has been neglected all too often, both nationally and at individual institutions. In order for the Nation's researchers to produce transformative science using advanced CyberInfrastructure, it will be essential to develop an interdisciplinary research computing workforce to support them. Such a workforce will depend heavily on personnel who are well trained in computational and data science; who have had exposure to programming paradigms, technologies and computational problems in a variety of fields; and who have compelling, exciting career paths open to them that can compete with traditional paths already available in academia and industry.

In our experience, researchers working in computation- or data-intensive fields have rarely received much formal training in advanced scientific computing or HPC. Often, much of their learning has been acquired informally from colleagues or a handful of well-thumbed "cookbooks," and it has usually focused on "getting stuff done," rather than developing a deep understanding of modeling, analytic, and computational methodologies. At the present time, we know of a good number of institutions (including Yale) that have begun to offer education in data science, but there seem to be very few that provide similar offerings in computational science. We believe that NSF and other funding agencies should address this situation by providing targeted funding for development of innovative training programs and curricula in both computational and data science that encompass both traditional academic coursework and, especially, short-courses, workshops, or on-line classes that will dramatically broaden accessibility. It seems particularly important to support programs that aim to equip students, post-docs, and others with the cross-disciplinary tools and skillsets needed to sustain lengthy careers in modern computational science that, over time, will undoubtedly involve work with researchers in multiple fields of science.

Matching CI Resources to Computational Science Research Needs:

National CyberInfrastructure facilities (such as those run by NSF and the Department of Energy) generally can provide only a fraction of the resources requested by their ever-growing communities of scientific research users. Even the most worthy projects often receive half or less of their well-justified requests. This poses real difficulties for research teams who have insufficient computational resources to successfully complete their research programs.

To clarify the main issue, consider a proposal for experimental investigation where the proposal asks for funds to buy key equipment for doing the research as well as the stipend of a graduate student to do the research. It would hardly be sensible to only fund the graduate student stipend but substantially cut the funds for the equipment. And yet, for many computational science proposals, the situation is very much like this: the proposals are funded based on the quality of the proposed science and broader impacts, but sufficient computational resources required to do the work are not allocated at the time of award or somehow guaranteed to be available over the life of the project.

This type of disconnect between the proposed research and the resources needed to do it leads to large uncertainty regarding final success. It would be very helpful if the NSF and other funding agencies were to plan for some type of direct connection between funded proposals, the scale of resources needed for the research, and the actual resources that can be guaranteed to the PI to do the funded work. This would require better coordination between the agencies and the national centers, as well as potentially changing what information must be provided by PIs and what will be funded in a successful proposal.